PROBLEMS OF ALGORITHMIC COMPOSITION OF SUBJECT INDEXES (BRIEF SURVEY OF FOREIGN LITERATURE),

SCIENTIFIC-TECHNICAL INFORMATION
(NAUCHNO-TEKHNICHESKAIA INFORMATSIIA),
No. 5, 1965, pp. 22-25

B. V. Iakushin

Translated from the Russian by
Joy B. Gazley and Wade B. Holland
The RAND Corporation

March 15, 1966

LT-65-102

RESUME

Major trends in developing methods of machine assignment of subject indexes are examined. Two approaches are discernible: a) the use of subject descriptors based on key words extracted from the document (usually from the title); b) the use of predetermined subject categories under which documents are indexed. It is noted that the systems that rely on statistical characteristics of the appearance of key words in the documents are the more promising.

The process of composing a subject index and the corresponding problems of its automation are divided into two parts: a) isolating basic concepts from the text of the document (subject categorization); b) correlating these concepts with the subject index headings (entry correlation).

Our primary interest is in algorithmic methods of solving these two problems; therefore, many types of indexes and their problems are not appropriate for consideration in this article.

The direction of experimental work in recent years in solving these problems has involved isolating significant (i.e., "key" or "basic") terms in a document. In practice, the problems of subject categorization and entry correlation are solved through the use of these sets of terms.

If one attempts to trace the development of these problems (in a logical more than a historical progression), the following are the major stages. Initially, attempts were made here to narrow the key-word concept.* In any category, a small and appropriately convenient number of key words was introduced. This was done by applying supplementary criteria, both formal and informal, relating to the principal parts of speech.

In several of the first experiments with key words, either all the derivatives of the words, or only the nouns and adjectives, were extracted [1]. Next, subclasses within these extracted classes were isolated according to various criteria. Criteria that won wide recognition were those that selected words that appeared with highest frequency in a text or in a class of texts and those that selected terms from document titles.

In solving the second problem, that of entry correlation, there were attempts at formulating indexes in which combinations of terms would be required for making entries, rather than separate terms. This results in narrower indexing and is better suited to practical problems of retrieval. In other words, it overcomes the difficulty that arises when key words are themselves used as index entries, wherein as soon as a key word is encountered in a document the document becomes correlated with that index entry (i.e., indexed under that heading).

---

*By "here" it is assumed the author refers to work done in the Soviet Union--Trans.

Because of the abundance of entries and the number of documents placed under any given heading, such an index is unsuitable for a large body of documents.

At present, the entry correlation problem is met in two ways:

1) The key word with the contextual phase in which it is found serves as the index entry;

2) Index entries (each with matching key words) are employed, but are not selected from text. By means of a thesaurus, key words in a text are matched against key words that are linked to index headings; as a result, the document is listed under the match-producing headings.

These methods of compiling subject indexes are based on the broad principles underlying information retrieval systems. Therefore, the organization of an index in any given subject field is often not subject-limited, but rather reduces to a method for handling particular logical, linguistic, and technical problems.

It is assumed that direct document retrieval is impractical. For many reasons, it is more convenient to use pre-established document descriptors, or retrieval tags, which characterize the subject of the document and its address in the storage file.

Descriptors that characterize a document's subject fields are nonunique, since each can apply to any document or to a whole set of documents. The information that is output about a particular document and its address comprises a unique descriptor, since it refers only to a single document.

Thus, there are two parts to the descriptor: the nonunique and the unique tags. The presence in the search descriptor of both parts is a necessary condition for retrieval from the file.

In order to better understand the significance of experiments in the field of automatic indexing, it follows to attempt to analyze a good index, one that enables retrieval of the necessary documents (or more accurately, of their unique descriptors) from a large data file, rapidly and without omissions or "noise."

The basic requirement of a subject categorization system is, apparently, that it generalize (apply to) the contents of several documents, rather than that it utilize nonunique descriptors that to some degree indicate the subject content of the documents (which is a problem of any descriptor system). In other words, a certain number of unique tags is assigned to each descriptor entry in the given file of documents. Any deviation from this number complicates the search.

Actually, if only one document is assigned to each subject descriptor (as occurs in many modern automatically-produced indexing systems), it is obvious that the number of descriptor entries will be larger than or equal to the number of documents. Systems of this sort are acceptable only for small files, since in large files the volume of subject entries would hinder the search. On the other hand, if the entries are so general that a large number of documents is listed under each, the system provides only a first-cut retrieval. For example, from a particular tabular-type file of 8000 documents, 1000 were extracted,

where a manual search would have resulted in selection of only 75 items.*

Thus we consider as a good system one from which in each category an optimum number of documents can be retrieved, the optimum number being determined on the basis of the total volume of the file.

All existing subject indexing algorithms reduce to a technique for identifying significant and key words in text; thus, the problem of subject categorization is resolved in a comparatively uniform manner. The problem of index entry correlation, as already noted, is solved in different ways: by directly utilizing the entries isolated from text (textual entries), or by assigning texts under pre-specified, manually-formulated categories. Proceeding from this difference, let us divide the systems under consideration into two classes:

1) Systems based on textually-derived subject categories;

2) Systems based on predetermined subject categories.

Following are short descriptions of what we consider to be typical algorithms for each class.

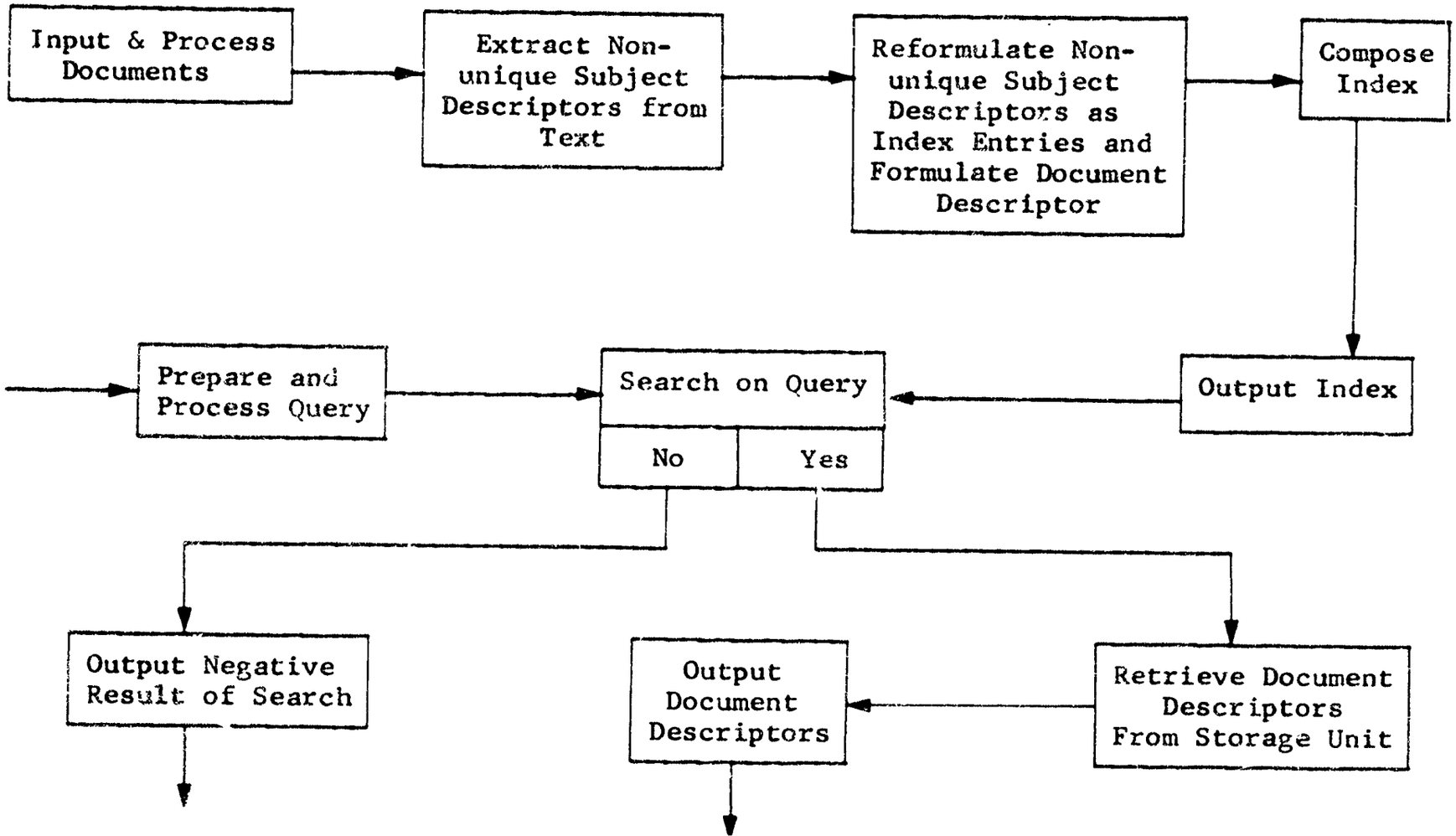
SYSTEMS USING TEXTUALLY-DERIVED SUBJECT CATEGORIES

An information retrieval system based on the use of a textually derived subject-categorization index can be represented by the following flow diagram.

---

*The original article contains a citation at this point of Ref. 4, p. 205. The citation is clearly in error, since there is no page 205 in Ref. 4 and since that article is not pertinent to the point being cited.--Trans.

BLOCK-DIAGRAM OF AN INDEX WITH TEXTUALLY-DERIVED SUBJECT ENTRIES



A simpler method of compiling an index for a journal is to introduce a list of key words in the computer and search for matching key words in the texts of the articles. Each key word is tied to a subject entry under which the document will be indexed if a match obtains. Such an experiment was undertaken in the laboratories of IBM (U.S.A.), and is described in Ref. 1. The results were admittedly unsuccessful because of the many subject entries and the large number of documents listed under each.

At the Institute of Interdisciplinary Sciences of the University of Pennsylvania, a subject index has been compiled in the form of a two-entry table.* Document numbers are entered in a wide left-hand column, while symbolic representations of terms describing the contents of the documents are denoted in columns to the right. For economy, several terms, represented by capital letters, are assigned to each column. The proper letter is entered into the cell at the intersection of its assigned column with the row containing the document number to which the symbolized key term applies. If several terms assigned to the same column are encountered in one document, the additional letters are indexed and entered into the nearest empty cell of the same column. During retrieval, a query term is sought down its proper column, while the other terms are sought by searching linearly along each row at which the query term has been found. If all the query

*At this point, the original Russian article cites Ref. 4; however, the citation is clearly in error since Ref. 4 pertains only to the work attributed to it below. None of the references at the end of the article appear to be correct for this citation.--Trans.

terms are found in the row, then the left column of the row is interrogated for the document number.

A similar index has been used for 30 years in the Patent Office of the Netherlands. At the present time, it is being converted to a punchcard system, because tabular organization of a large file becomes too unwieldy to be conveniently used. The growth of the dictionary, inevitable in an algorithmic subject-categorization system, increases the size of the table even more, further complicating retrieval. Moreover, index systems based on non-ordered collections of subject entries result in a great deal of "noise" during search, caused by improper combinations of terms.

The most widely used systems (at least in the U.S.A.) for indexing small periodical publications are the "key-word-in-context" (KWIC) indexes, where the title of the article is a nonunique descriptor (index entry). Titles are arranged alphabetically according to their key words. A key word dictionary is first fed into the computer; if one of the key words is encountered in a title, the title is entered in the index under that key word. The actual entry consists of the article title preceded by the applicable key word, or the article title with the applicable key word specially indicated in some fashion. The same title is entered as many times as there are key terms in it. Thus, there are more entries than documents and the index is relatively large.

In order to keep the index from becoming unwieldy, encoded unique descriptors (alphabetic) are employed in the file in place of full bibliographic citations. These

letter codes are located in the file in alphabetic order, and are assigned to the articles so as to constitute unique descriptors.

Examples of KWIC applications include the "Chemical Titles" system that indexes the journal Chemical Abstracts, the "Biochemical Title Index" which is similar to the "Permuted Titles" index to the Bell Laboratories Record (described in Ref. 3), and others.

The shortcoming of such systems is that they are suitable only for publications with a limited number of articles; a no less serious deficiency is the fact that only the title, rather than the entire text, is used as the basis for the nonunique descriptors. Because titles are so often void of real content (especially concerning questions of theory and technological history), retrieval of needed documents is limited. In our opinion, these experiments involving lexical comparisons of index entries against document titles demonstrate these conclusions. If the entry under which a document is indexed and its title do correspond to the actual contents of the document, then, naturally, a correlation can be assumed. Additionally, it should be recognized that a greater volume of common terms is utilized in an index and in titles than in running text.

Data on correlation between titles and index headings are presented in Ref. 5. Of all possible relationships (identical words, inflectional variants, synonyms, synonym-inclusions--"tranquilizing agent" and "tranquilizer"), for the Index Medicus the synonym-inclusions alone achieved an 86-percent correlation; the range was 19 to 45 percent

for the Index-Handbook of Cardiovascular Agents, and 13 to 39 percent for the U.S. National Institutes of Health [Research Grants] Index for 1961. Although these data do not permit direct conclusions concerning disparity between title and text, they do, however, testify to the complexity of the terminological relationships between them.

Aware of this insufficiency, H. P. Luhn used all text sentences in the original KWIC system, retainin only the highest-frequency words as key terms. A text phrase entered into his index would be one consisting only of significant terms, including the maximum-frequency term. Preference, obviously, is given to phrases found in the title or nearest to it, rather than to phrases from elsewhere in the article containing the same high-frequency term. In the index, the entered phrase is divided into three parts: the words standing to the left of the key term; the key term; and the words standing to the right of the key term. The phrases (or entries) are ordered alphabetically in the index according to the key term.

Although this method permits more accurate processing of texts with meaningless titles, its principle of keyword frequency has not found wide acceptance, apparently because of a reluctance to thus complicate the index-compilation algorithm.

SYSTEMS USING PRE-ASSIGNED SUBJECT CATEGORIES

In the methods described, the index subject categories isolated from text have not been of a constant and generalized character. Generality is extremely important for

large files of data, as noted above. Therefore, resolving this specific problem is of paramount importance to some systems. A diagram of a system employing pre-assigned entries is shown below.

Such an index assumes the following specific developmental steps:

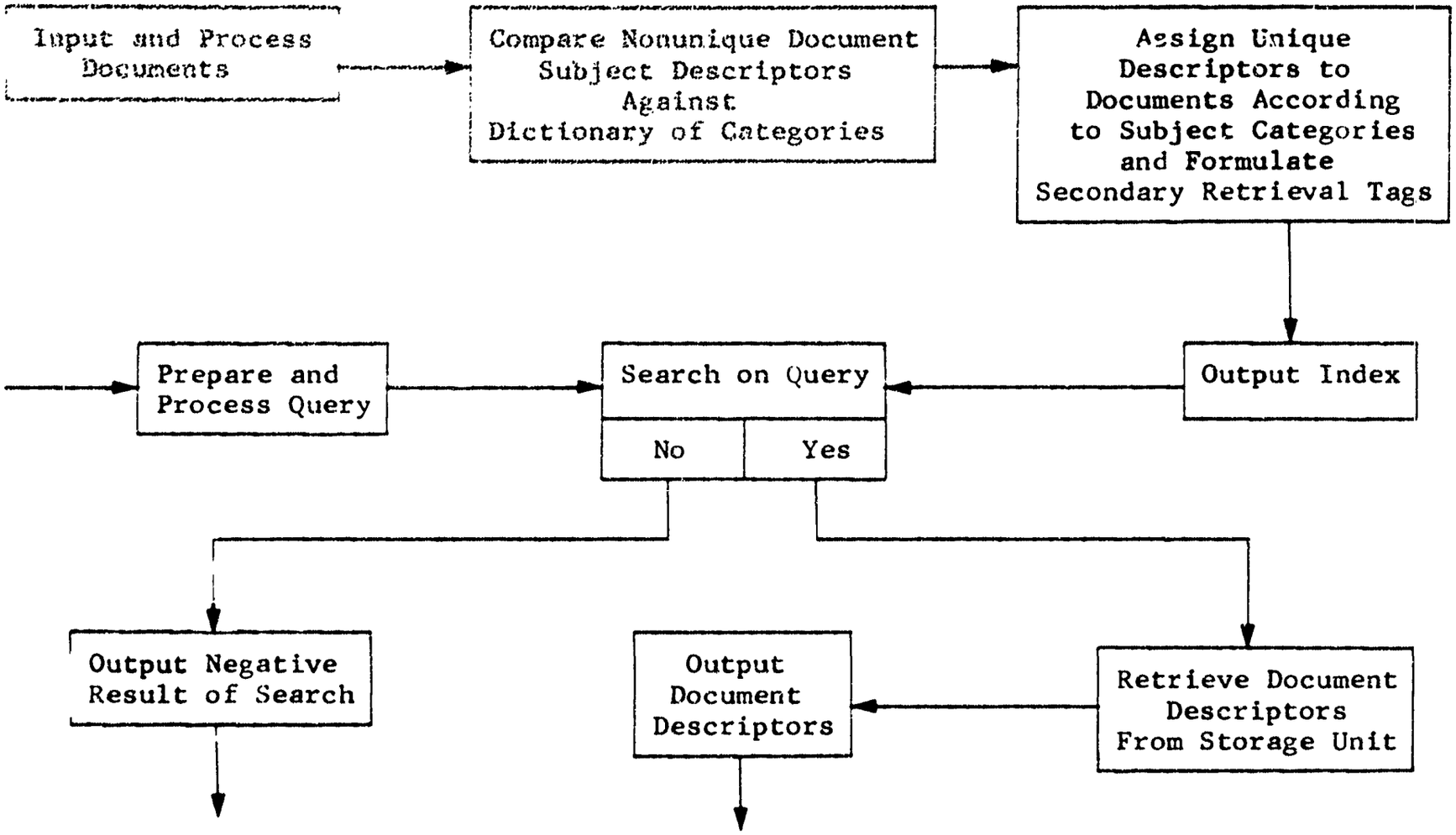
1) Manual compilation of a fixed list of subject categories;
2) Analysis of text terminology by classes of documents, corresponding to the subject categories;
3) Compilation of a list (glossary) of significant or key terms with a notation of the properties of each (usually statistical properties for each document class);
4) Formation of sets of terms extracted from texts of documents as nonunique search tags for the documents;
5) Comparison of these sets against the dictionary of subject categories;
6) Assignment of unique descriptors to indexed documents.

It should be noted that, in these systems, nonunique document tags based on a document's subject matter (primary descriptors) are also replaced by nonunique tags assigned in the subject category index (secondary descriptors).

The annual index developed for the abstracting journal Federation Proceedings of American Societies for Experimental Biology is a variation of the pre-assigned categories' index [4].

The set of significant terms, the bibliographic citation, and the document number are input to the computer. Beforehand, a subject-category list with a set of terms for

BLOCK-DIAGRAM OF AN INDEX WITH PRE-ASSIGNED SUBJECT ENTRIES



each category (i.e., sub-headings) is input. These terms become the key words for texts indexed under that heading.

Terms found in the document are re-edited using a machine-stored synonym dictionary (a thesaurus): synonyms are reduced to a single form--i.e., key words in the subject category listing are substituted. There is an algorithm (not described in Ref. 4) for comparing the set of significant terms from the document against the subject-category entries and for entering the document under the corresponding category. The page number and location of the abstract in the Federation Proceedings constitute the article's unique descriptor. The experiment has passed into the production stage.

A different system, based on statistical characteristics of terms, is found in Ref. 6. The highest-frequency key terms for a document under any entry are isolated from all the significant terms found in the text. On the basis of an analysis of 260 texts, 90 key words were isolated from 1000 significant words. Under each subject category, a dictionary was set up of key terms with their "attribute numbers," representing the probability $P(C_i/r_j)$ that a document belonging to category r_j contains key term C_i.

$$P(C_i/r_j) = \frac{n}{m} ,$$

where n is the number of occurrences of the term C_i in documents indexed under category r_j, and m is the number of occurrences of all key terms in the given category.

Then the probability that a document containing key terms $C_e, C_m, \ldots, C_s$ belongs to category r_j is

$$P(r_j/C_e,\ldots,C_s)=K\cdot P(r_j)\cdot P(C_e/r_j)\cdot P(C_m/r_j)\ldots P(C_s/r_j),$$

where K is a scale factor, and $P(r_j)$ is the derived probability that the text belongs to category r_j.

$$P(r_j) = \frac{l}{s},$$

where l is the number of documents listed under category r_j by a human indexer, and s is the number of documents in the file.

Applying the algorithm in processing a file, it was shown that, based on statistical characteristics of 260 documents and 32 categories, machine indexing provided accurate assignment of documents to subject categories in 85 percent of the cases.

The basic deficiency of all library classification schemes, the impossibility of formulating an indexing (or subject-categorization) algorithm that facilitates the use of a strictly hierarchical classification system in machine retrieval, can be overcome by employing approaches of this sort. Experiments in designing such types of systems are becoming more widespread [7,8].

The labor-consuming nature of their design and, even more so, of compiling a glossary for each subject category, are the principal drawbacks to using statistical methods of subject indexing. However, in our opinion, these methods in particular will dominate the immediate future.

It would seem reasonable to combine them with KWIC-type systems of index compilation; if the latter can, with relative accuracy, describe the contents of a document for subject determination, then the former can insure the generalization necessary for larger files of documents.

CONCLUSIONS

The basic purpose of a subject index, accurate and succinct characterization of subject matter, is best realized in systems using textually-derived entries. But this is done on the level of vocabulary. A subject topic is considered to be characterized if either individual words or a phrase extracted from the text can be regarded as an integral lexicographical unit. Figuratively speaking, we are still obliged to judge an automobile by its wheel.

In order to characterize subject matter, it is necessary to take not just one part of a text, but rather the whole text and extract from it the overall content. In a large file of documents, the overall-content entry cannot be exclusive to just one text, but must apply to a series of documents. This list of documents must be neither too long nor excessively short.

The process of assigning documents according to subject entries in systems using pre-assigned categories is also accomplished at the lexical level. A semantic correspondence between the contents of a document and a subject category is realized on the basis of comparing logically and grammatically unordered sets of terms. Less direct connections between the subject entry and the

document contents reduce the specificity of its description for retrieval.

The formulation of a nonunique descriptor of overall content (i.e., a search tag) for a text can be handled at the level of logical and syntactical analysis and synthesis. Only by working out the relationships between the parts of the text contents and by making the transition from singular relationships and textual units to relationships based on a rich, well-developed thesaurus will deeper analysis of text be possible. However, all this can be of no value if methods of formalizing logical relationships between units of text are not developed. The most realistic approaches to this end involve earnest application of syntactical analysis of a sentence; then its translation to logical analyses first of the sentence and then of the text as a whole; and, finally, logical reorganization of the text according to its subject matter.

Received, November 12, 1964

REFERENCES*

1. Baxendale, P. B., "Machine-Made Index for Technical Literature. An Experiment," IBM J., Vol. 2, No. 4, 1958, pp. 354-361.
2. Luhn, H., "Key-Word-in-Content Index for Technical Literature," Amer. Docum., Vol. 11, No. 4, 1960, pp. 288-295.
3. Bell Lab. Rec., Vol. 39, No. 6, June 1961, p. 211 (news item).
4. Schultz, C. K., and P. A. Schwartz, "A Generalized Computer Method for Index Production," Amer. Docum., Vol. 13, No. 4, 1962, pp. 420-432.
5. O'Conner, John, "Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems," Amer. Docum., Vol. 15, No. 2, pp. 96-104.
6. Maron, M. E., "Automatic Indexing: An Experimental Inquiry," J. ACM, Vol. 8, No. 3, 1961, pp. 404-417.
7. Williams, John H., Jr., "A Discriminant Method for Automatically Classifying Documents," AFIPS Conference Proceedings, Vol. 24, Spartan Books, Inc., Baltimore, Maryland; Cleaver-Hume Press, London, 1963, pp. 161-166.
8. Borko, Harold, and Myrna Bernick, "Automatic Document Classification," J. ACM, Vol. 10, No. 2, 1963, pp. 151-162.

*Errors in reference citations in the original Russian article have been corrected.--Trans.